Data Analysis and Hypothesis Testing Using the Python ecosystem

Hypothesis testing

Stavros Demetriadis
Assc. Prof., School of Informatics,
Aristotle University of Thessaloniki

sdemetri@csd.auth.gr
http://mlab.csd.auth.gr/sdemetri

Overview

- Becoming productive with the **Python ecosystem**

- **Hypothesis testing:** An exemplary case analysis
  - *Introduction: Key ideas*
    - Research design & Experimental design
    - Reliability, Validity, Generalizability
  - *The meaning of hypothesis testing*

Becoming productive with the **Python ecosystem**

- **Productive = Problem solving**

Data representation and coding

Using domain-specific software to solve problems

pandas
panel data

Numpy
numerical Python

Basic Python

Roots: Python class hierarchy

1) Use **Scipy**
Scipy = Scientific Python

2) Use other **python-related** software packages

3) Write **your own** domain-specific code

Contribute to community

Hypothesis testing

An exemplary case analysis

Introduction: Key ideas

Research design

- **Research design** is a broad term referring to **the work that researchers do when planning a research activity**.
- A research design may take **various forms** depending on…
  - … the domain, the objective of the research, the type of available data and the understanding one seeks to elicit.

- **Major research design paradigms**
- a) ***Quantitative***: measuring *quantified variables* and reach conclusions by processing quantitative data (e.g. testing hypotheses, implementing predictive modeling techniques, etc.)
- b) ***Qualitative***: collecting and analyzing *qualitative data*, that is, non-measurable constructs (e.g. opinions and artifacts) that help elicit a deeper understanding and better interpretation
- c) ***Mixed methods***: apply *both quantitative and qualitative analysis* in a complementary fashion to get the best out of the two worlds (triangulation!)

Three steps in quality research design

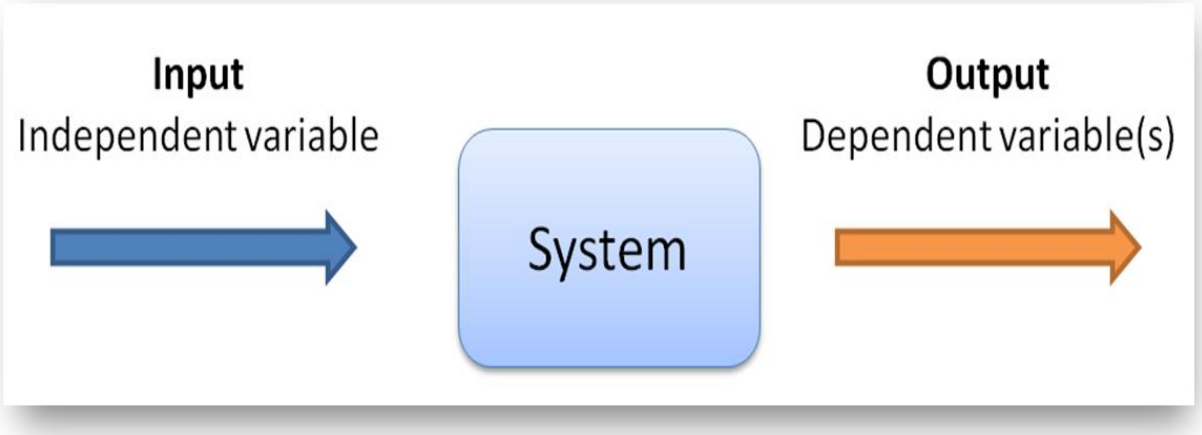
- **(a) Applicability of the method**
- *whether a specific research method is appropriate to apply in a specific research situation*.

- **(b) Implementation details**
- *sequencing implementation actions and operations, managing any possible peculiarities and "watch out for" points, and skillful use of necessary tools*.

- **(c) Reporting**
- *Reporting the implementation and results of a research method*

Quantitative paradigm: **Experimental design**

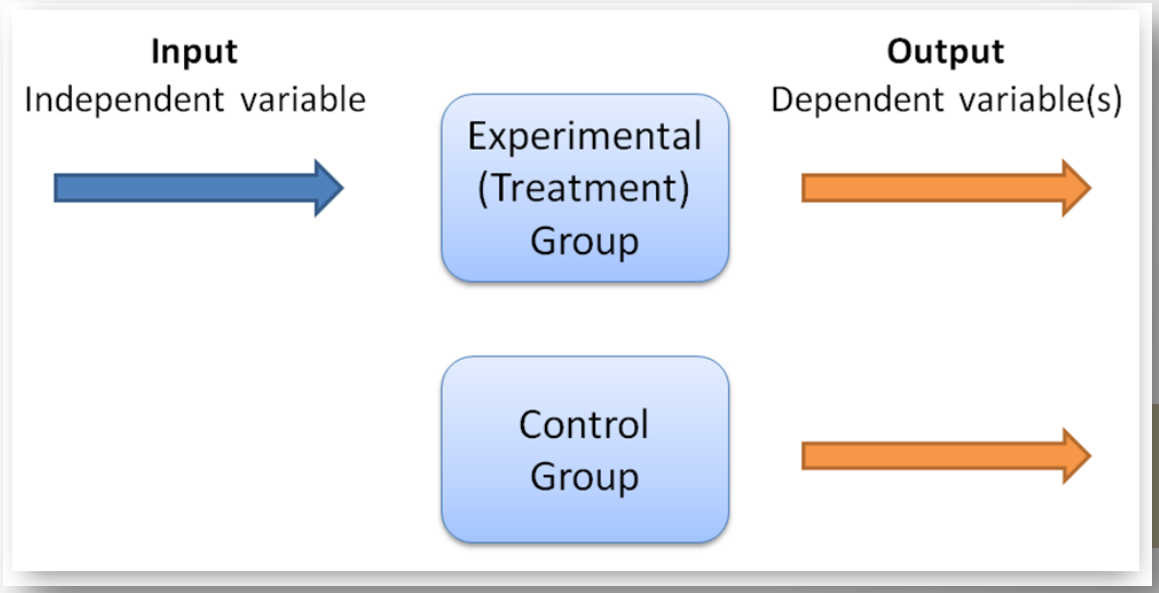
- **Experimental** design is the type of research activity typically implemented **in the lab** where controlling independent variables is feasible.
- The researcher *maintains control* over multiple environmental interfering factors
- Studies the impact of *one independent variable* on the behavior of the system (the dependent variables) under study.



A typical design: **Experimental vs. Control group**

- **Experimental group** ('treatment' group): the group where the independent factor (variable) has an impact
  - *patients* given an experimental drug or *students* learning with an innovative method/technology
- **Control group:** the group where no special treatment is applied
  - *patients* are given a placebo drug or *students* learn as usually without any innovation

R	X	O
R		O



Sampling

- **Sampling:** the process of group formation by *selecting members* from the population that we would like to investigate
- **Random** sampling: selected *completely randomly* from population. The simplest non-biased approach to getting a sample.
- **Stratified** sampling: when a sample is *divided into subgroups* (for example, boys and girls) as in the original population
- **Convenience** sampling: getting *convenient* (easy to find and use) sample groups
  - For example, two student intact classes used as treatment and control groups

Measuring

- **Measuring**: the process of using a **measuring instrument** to get numerical data (numbers) representing the value of the **measured construct**.
- A '**measured construct**' is a *conceptual construct* (a concept) referring to the property that we aim to measure.
  - For example: '*individual learning*'
- A '**measuring instrument**' is a tool (pre-existing or ad hoc developed) for measuring the construct.
  - For example: a *questionnaire* for measuring individual learning

Key issues when measuring

- **Reliability**: how reliable are the measurements?
- **Validity**: are the measuring instrument(s) valid?
- **Generalizability**: after analyzing data can conclusions be generalized?

Reliability

- **Reliability** refers to *how consistent or repeatable is the measurement responding to changing conditions but without being affected by irrelevant factors*.
  - *Consistency*: A questionnaire to measure individual learning should provide similar outcomes when used with similar samples (groups of students from the same population)
  - *Irrelevant factor*: The questionnaire outcomes should not be affected (for example) by the outside temperature

- **Reliability** in statistics and psychometrics is the **overall consistency** of a measure (Wikipedia)

Reliability

Reliable and **Valid** instruments is what we need

An instrument can be reliable without being valid

An unreliable (hence not valid) instrument is never useful

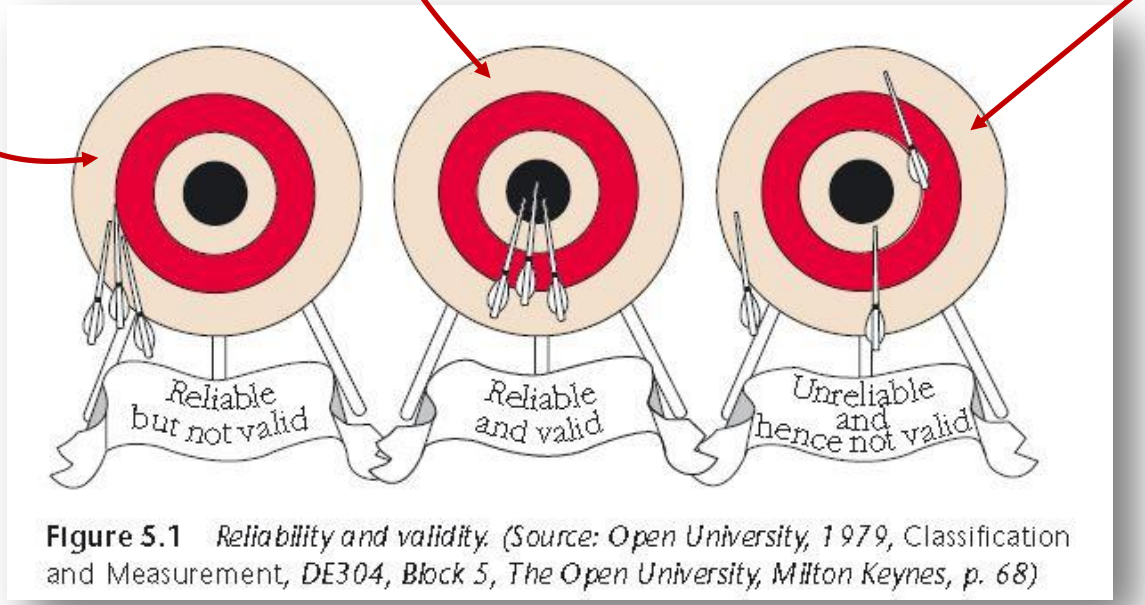


Figure 5.1 *Reliability and validity. (Source: Open University, 1979,* Classification and Measurement, *DE304, Block 5, The Open University, Milton Keynes, p. 68)*

Reliability 3/3

- Various types of reliability:
  - **Inter-rater**: the degree of agreement among raters (wikipedia)
  - **Test-retest:** is the variation in measurements taken by a single person/instrument under the same conditions, in a short period of time (wikipedia)
- Common reliability measure: Cronbach's alpha
  - Measure of internal consistency: how closely related a set of items are as a group (SPSS FAQ, Univ. of Virginia)
  - High reliability: **a >= 0.7**
- Reliability of measuring instruments ➔ high probability of getting *similar results under similar conditions*

Validity

- **Validity** generally refers to *the quality of being correct and accepted by experts and authorities*.
- **Being valid** is the property of an instrument *truly measuring* what their developers claim it does

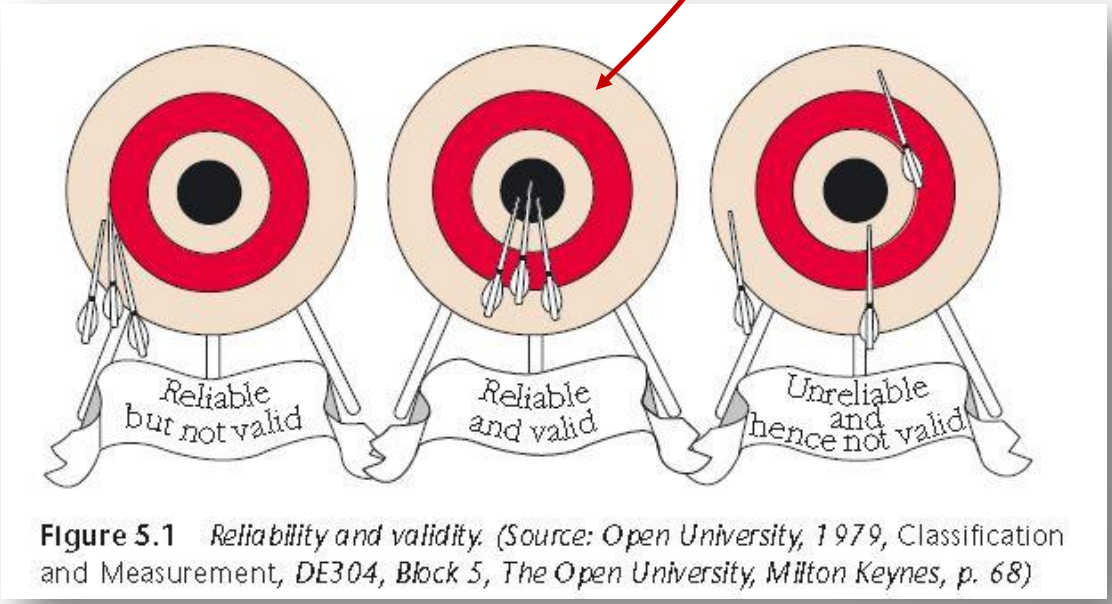


Figure 5.1 *Reliability and validity. (Source: Open University, 1979,* Classification and Measurement, *DE304, Block 5, The Open University, Milton Keynes, p. 68)*

Validity 2/2

- **Construct validity** of design: *whether the design is operationalized in a way that fits well to the theoretical underpinnings of the study*
  - for example, if the theory predicts an impact on 'group learning' but the research design measures only the individual learning, then this is *a threat to construct validity*

- **Internal validity** of design: *are the study conclusions regarding cause-effect valid (true)?* (when trying to establish a causal relationship)
  - for example, if the researcher erroneously concludes that independent factor X has an impact on dependent variable Y, then this is *a threat to internal validity*

Generalizability

- **Generalizability:** *the extent to which the conclusions of a specific study can be generalized over a broader population*
  - For example: if a study on students' language skills has recruited only English speaking students then are conclusions generalizable over student population speaking different languages?
- Many ways for generalizing
  - from one setting to a larger number of settings?
  - from one rater to a larger number of raters?
  - from one set of items to a larger set of items?
- Further reading: Generalizability theory (Wikipedia)

The meaning of hypothesis testing

Null hypothesis

- A **hypothesis** (plural: hypotheses; from Greek 'υπόθεσις') *is a specific statement of prediction relevant to the phenomenon under study*.

- For example, suppose that your research question is: *Does background music in a multimedia learning environment have a positive/negative impact on students who use this environment to learn?*

- **Null hypothesis** (H_0): "Background music has no impact whatsoever on students' learning"

Hypothesis directionality

- '**non-directional**' hypothesis: *does not predict* the direction of impact (positive/negative).
  - "*Background music has no impact whatsoever on students' learning*"

- '**directional**' hypothesis: *it predicts* the direction of impact (positive/negative).
  - "*Background music has a positive impact on students' learning*"

What is hypothesis testing

- **Hypothesis testing** refers to the procedure of *applying appropriate statistical controls that help formulate conclusions* regarding the "truthfulness" of the stated hypotheses.

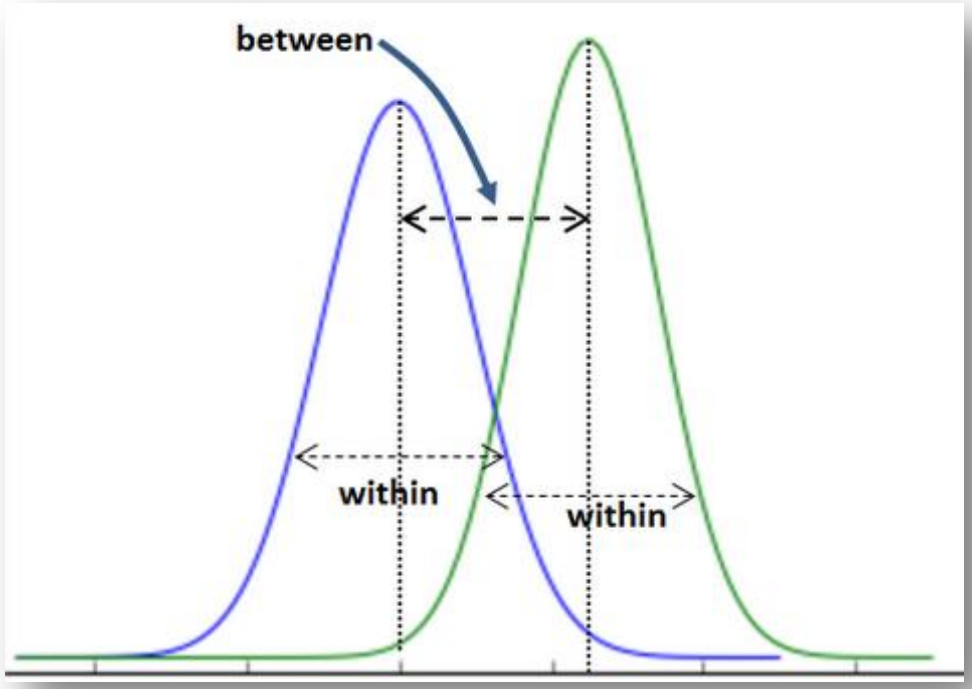
- **Hypothesis tests** (or 'statistical controls') are *a whole range of statistical algorithms* for data processing that return:
- (a) a **value for the 'test statistic'** we need to compute, and
- (b) the **probability** for this value to appear

An example 1/2

- Research design:
  - **Experimental** group: N_E = 25, study **with** background music
  - **Control** group: N_C = 22, **no** background music during study
  - Learning instrument: knowledge questionnaire (a 0-100 scale),
  - Two conditions, Post-test only study
- Outcomes:
  - E group **performance**: M = 75, sd = 8.7
  - C group **performance**: M = 65, sd = 9.5
- Statistical control: T-test
  - **Statistic**: t
  - **Probability**: p

An example

- Experimental group
  - Control group

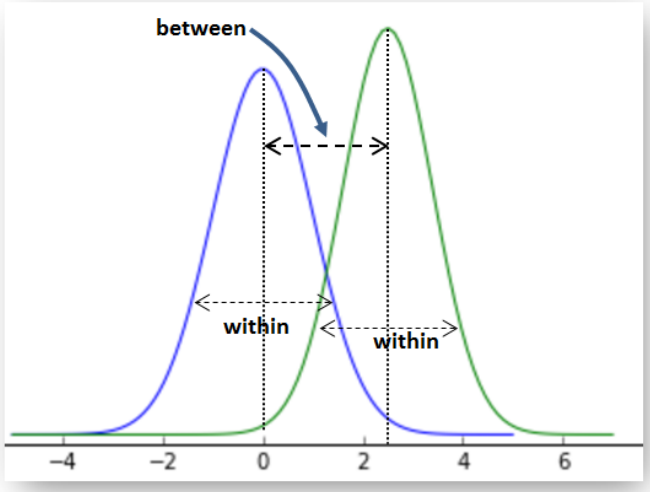


- Is the observed difference in mean values of the two groups:
- A) a random fluctuation, or
- B) due to impact of the independent factor (i.e. background music)

The rationale for hypothesis testing

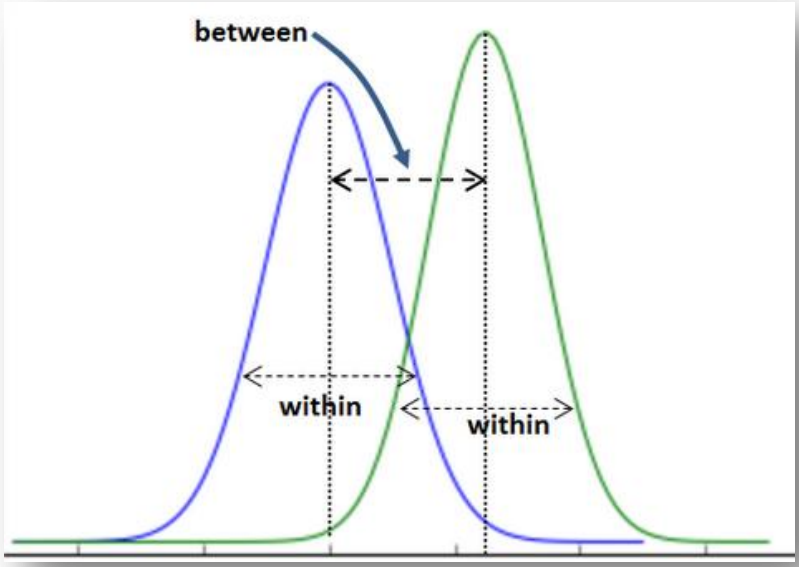
- All statistical hypothesis tests estimate a **'signal to noise'** ratio.
- *Nominator*: '**Signal**' = between groups variability (unexpected)
- *Denominator*: '**Noise**' = within groups variability (expected)

$$\text{measure} = \frac{\text{signal}}{\text{noise}} = \frac{\textbf{Between} \text{ groups variability}}{\textbf{Within} \text{ groups variability}}$$



- If the ratio is **too big** then something significant is going on
- But how big is 'too big'?

Check ‘p’ against the ‘a’ threshold



- Suppose t-test control gives the following outcomes:
- **t = 3.7, p = 0.038**

- Set a ‘significance’ threshold:
- Usually: **a = 0.05**

- Hence: **p < a**
- Conclusion: The difference between the two groups is **statistically significant** ➔ the independent factor (b.m.) had indeed an impact

Based on this probability we either:

- ***'reject***' the null hypothesis and support an 'alternative' hypothesis that better interprets the data
  - For example: *Background music has a significant positive impact on student learning*

- ***'fail to reject***' the null hypothesis in which case we stay with it,
  - until -of course- new data may lead us to reject it
  - note that, in general, we avoid the expression "to accept the null hypothesis"

But how do we estimate ‘p’ value?

- **Variable Distributions**!

- We need to know the *variable distribution* of the statistic that we are using (t, F, etc.)
- Then we can *get the probability p* that the variate lies within a specific range of values
- We can answer questions like: *what is the probability that the statistic (t, F, etc.) has a value equal to the one computed based on the available data?*

Summarizing

- Hypothesis testing:
- a) Defines a **statistic** (t-test, F-test, etc.) in the form of a "**signal/noise" ratio** and provides a means (algorithm) for its computation
- b) Uses a **measure-relevant distribution** to compute the **probability** that a measure lies within a range of values
- C) Concludes:
  - If p <= a then **reject** H_0
  - If p> a then **fail to reject** H_0